Welcome to IMS

This is the website for Introduction to Modern Statistics, First Edition by Mine Çetinkaya-Rundel and Johanna Hardin. Introduction to Modern Statistics, which we’ll refer to as IMS going forward, is a textbook from the OpenIntro project.

The book will always be available for free here. It is also available in PDF (for free or for the amount you choose to donate to the OpenIntro project) on Leanpub and in black & white paperback for purchase for $20.

Download PDF
Purchase paperback

Copyright © 2021.

First Edition.

Version date: November 06, 2021.

This textbook and its supplements, including slides, labs, and interactive tutorials, may be downloaded for free at openintro.org/book/ims.

This textbook is a derivative of OpenIntro Statistics 4th Edition and Introduction to
Poll

How do you identify?

a. mathematician
b. statistician
c. computer scientist
d. something else
Poll

How often do you teach intro stats?

a. every semester
b. sometimes
c. I’ve taught it a few times in my career
d. not yet (still waiting for my chance!)
Start with graphics

Then on to modeling

And wrap up with inference

... hopefully leaving them wanting to learn more!
Visualization is…

a. a big part of what I teach in intro stats
b. part of just-in-time teaching (e.g., need a boxplot with ANOVA)
c. emphasized only as a method of descriptive statistics
d. not something I cover in intro stats
Data visualization

Start with bivariate visualizations
Data visualization

Quickly move on to multivariate visualizations
Introduce “non-standard” visualizations

Data visualization
Use visualizations beyond just data exploration

... but more on that in a bit!
Start with graphics

Then on to modeling

And wrap up with inference
Modeling

As a way to summarize relationships

- Between two numerical variables: simple linear regression
- Between one numerical and many numerical and categorical variables: multiple linear regression
- Between a categorical and many numerical and categorical variables: logistic regression
Questions?
Start with graphics

Then on to modeling

And wrap up with inference
Modern Inference

- Hypothesis testing
  - Randomization tests vs. CLT approaches (t-test, z-test of proportion, ANOVA, chi-squared)

- Confidence intervals
  - Bootstrapping vs. CLT approaches (t-intervals, z-intervals for proportions)

- Model validation
  - Cross validation vs. p-values to assess variables in multivariable models
Chp 17: 2 proportions

17.1 Randomization test for the difference in proportions
17.1.1 Observed data
17.1.2 Variability of the statistic
17.1.3 Observed statistic vs null statistics
17.2 Bootstrap confidence interval for the difference in proportions
17.2.1 Observed data
17.2.2 Variability of the difference in sample proportions
17.2.3 Bootstrap percentile vs. SE confidence intervals
17.2.4 What does 95% mean?

17.3 Mathematical model for the difference in proportions
17.3.1 Variability of the statistic
17.3.2 Confidence interval for the difference between two proportions
17.3.3 Hypothesis test for the difference between two proportions

Chp 22: many means

22.1 Case study: Batting
22.2 Randomization test for comparing many means
22.2.1 Observed data
22.2.2 Variability of the statistic
22.2.3 Observed statistic vs. null statistic
22.3 Mathematical model for test for comparing many means
22.3.1 Variability of the statistic
22.3.2 Observed statistic vs. null statistics
22.3.3 Reading an ANOVA table from software
22.3.4 Conditions for an ANOVA analysis
Hypothesis testing

Randomization test

The test statistic for comparing two means is a $T$.

The $T$ score is a ratio of how the groups differ as compared to how the observations within a group vary.

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

When the null hypothesis is true and the conditions are met, $T$ has a $t$-distribution with $df = \min(n_1 - 1, n_2 - 1)$.

Conditions:
- Independent observations within and between groups.
- Large samples and no extreme outliers.

$t$-test
Intervals

Bootstrap intervals

Margin of error for $\bar{x}_1 - \bar{x}_2$.

The margin of error is $t^*_{df} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ where $t^*_{df}$ is calculated from a specified percentile on the t-distribution with $df$ degrees of freedom.

t-intervals
Modeling

\[ E[body_{mass,g}] = \beta_0 + \beta_1 \times bill\_length\_mm \]

\[ body\_mass\_g = 362.31 + 87.42 \times bill\_length\_mm \]

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>362.3</td>
<td>283.4</td>
<td>1.28</td>
<td>0.2019</td>
</tr>
<tr>
<td>bill_length_mm</td>
<td>87.4</td>
<td>6.4</td>
<td>13.65</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

p-values

Cross validation
Benefits to the parallel modern approach

● Modern methods capture underlying inference structure
  ○ What is a sampling distribution?
  ○ What does it mean to make claims about a population based on a sample?
● Students are conversant in methods which are being used more and more frequently in industry.
● Students are forced to consider: which method is better?
Ideas for a classroom discussion

● What types of data structure (distribution, variability, outliers, sample size) "should" the test be able to assess?

● What would we hope to see if we were focused on "good" Type I error rates?

● What would we hope to see if we were focused on "good" Type II error rates?

● Which of the considerations above seems most important? Data structure? Type I error? Power?

● What makes a confidence interval method "good"?
Disadvantages to the modern approach

- Most of science continues to lean heavily on methods derived from the central limit theorem.
- Software is often set-up to easily apply CLT methods (although that software is typically more opaque than code which runs simulation based methods, e.g., the infer R package, infer.tidymodels.org)
Using computational / modern methods is

a. something I was taught and/or feel comfortable teaching
b. something I’m not comfortable with but am excited by and want to incorporate into my teaching
c. something that I’m not comfortable with introducing to a classroom
d. something I’m seeing today for the first time

Poll
Beyond IMS

synergy with other freely available resources

- excellent applets exist
  - Rossman/Chance
  - StatKey
- tactile in-class activities
Questions?
Welcome to IMS

This is the website for Introduction to Modern Statistics, First Edition by Mine Çetinkaya-Rundel and Johanna Hardin. Introduction to Modern Statistics, which we’ll refer to as IMS going forward, is a textbook from the OpenIntro project.

The book will always be available for free here. It is also available in PDF (for free or for the amount you choose to donate to the OpenIntro project) on Leanpub and in black & white paperback for purchase for $20.

Download PDF
Purchase paperback

Copyright © 2021.

First Edition.

Version date: November 06, 2021.

This textbook and its supplements, including slides, labs, and interactive tutorials, may be downloaded for free at openintro.org/book/ims.

This textbook is a derivative of OpenIntro Statistics 4th Edition and Introduction to
All the ways to get the book

- **HTML textbook freely available**
- **PDF** also **freely available** — distributed through LeanPub with a suggested donation (min $0)
- **Paperback** available at **low cost** — cost of printing + minimal royalty that goes back to OpenIntro (US-based nonprofit)
Pedagogy

- Applications as motivation
- Real, recent, relatable datasets
- Emphasis on data exploration, multivariable relationships, and statistical reasoning
- Lots of end of chapter exercises, with solutions to odd numbered questions in the back
- Guided practices and worked examples interspersed in text

**Example:**

We calculated a slope coefficient of 0.74 for \textit{bankruptcy} in Section 8.1 while the coefficient is 0.386 here. Why is there a difference between the coefficient values between the models with single and multiple predictors?

If we examined the data carefully, we would see that some predictors are correlated. For instance, when we modeled the relationship of the outcome \textit{interest\_rate} and predictor \textit{bankruptcy} using linear regression, we were unable to control for other variables like whether the borrower had their income verified, the borrower’s debt-to-income ratio, and other variables. That original model was constructed in a vacuum and did not consider the full context of everything that is considered when an interest rate is decided. When we include all of the variables, underlying and unintentional bias that was missed by not including these other variables is reduced or eliminated. Of course, bias can still exist from other confounding variables.
Poll

What software do you use?

a. R
b. python
c. other statistical software
d. applets / web interfaces
e. I don’t use any software
Interactive tutorials in the browser

- Linked from the textbook
- Step-by-step instructions
- No software installation
Lesson 1 - Bootstrapping for estimating a parameter

Welcome

Rent in Manhattan

Review: Percentile and standard error methods

Next is a two part exercise: First, generate 15000 bootstrap distributions of rent in the manhattan data frame and record the median of each bootstrap distribution.

- Specify that rent is the response variable.
- Generate 15000 bootstrap replicates.
- Calculate the median of each distribution.

---

Code

1 # Generate bootstrap distribution of medians
2 rent_med_cl <- manhattan %>%
3   # Specify the variable of interest
4   ___(response = ___) %>%
5   # Generate 10000 bootstrap samples
6   ___(reps = ___ + type = "__") %>%
7   # Calculate the median of each bootstrap sample
8   ___(stat = "__")
9
10 # Take a peek
11 glimpse(rent_med_cl)
Computing Labs

- Linked from the textbook
- Step-by-step instructions, but a bit less hands on than tutorials
- Designed for R and RStudio
Computing

R packages containing all datasets used in OpenIntro materials

Depends:

R (&ge; 2.10), airports, cherryblossom, usdata
Data sets

Another way to access all datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>golub.train</td>
<td></td>
<td></td>
</tr>
<tr>
<td>goog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gsv.poll</td>
<td></td>
<td></td>
</tr>
<tr>
<td>govrace10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gpa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gpa_iq</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gpa_study_hours</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gradistv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gsearch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gas2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>health_coverage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>healthcare_law_survey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>heart_transplant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>helium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>helmet</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
License

Attribution-ShareAlike 3.0 United States
(CC BY-SA 3.0 US)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.
Poll

Have you used an open source resource for your class?

a. yes
b. no, but I want to
c. no
Teacher Resources

We provide access to additional resources to help support classroom instruction, testing, and more.

Get Verified

Most resources on this page are restricted to Verified Teachers. If you're a teacher and would like to access these resources, please start by getting Verified.

Apply for Teacher Verification
Most teachers are verified within 3 business days

Learn about Teacher Verification
Benefits, options to apply, and the verification process

Essentials

Invite another teacher to OpenIntro
Available to Verified Teachers

Request a textbook desk copy (US only)
Available to Verified Teachers, click here to register

Get involved with OpenIntro
We have several projects where we could use your help

Solutions

Exercise solutions are available for each of the following textbooks. Solutions cannot be shared except in the limited capacity described in the PDFs. For fellow teachers who want access to the solutions, please invite them to OpenIntro!
openintro.org/teachers/get_involved

- **Datasets**: Data hunter
- **Exams**: Exam contributor
- **Exercises**:
  - Exercise contributor
  - MyOpenMath exercise conversion
- **Labs**:
  - Lab solution developer
  - New software owner
  - New lab developer
- **Modules**:
  - Module developer
  - Module editor
- **Translation**
- **Your ideas?**
thank you!

bit.ly/ims-amatyc